Biophysical Letter

Statistical Inference for Nanopore Sequencing with a Biased Random Walk Model

Kevin J. Emmett,[1,*] Jacob K. Rosenstein,[5] Jan-Willem van de Meent,[2] Ken L. Shepard,[4] and Chris H. Wiggins[3]

[1]Department of Physics, [2]Department of Statistics, [3]Department of Applied Physics and Applied Math, and [4]Department of Electrical Engineering, Columbia University, New York, New York; and [5]School of Engineering, Brown University, Providence, Rhode Island

ABSTRACT Nanopore sequencing promises long read-lengths and single-molecule resolution, but the stochastic motion of the DNA molecule inside the pore is, as of this writing, a barrier to high accuracy reads. We develop a method of statistical inference that explicitly accounts for this error, and demonstrate that high accuracy (>99%) sequence inference is feasible even under highly diffusive motion by using a hidden Markov model to jointly analyze multiple stochastic reads. Using this model, we place bounds on achievable inference accuracy under a range of experimental parameters.

Received for publication 18 February 2015 and in final form 10 March 2015.

*Correspondence: kje@phys.columbia.edu

Rapid advances in DNA sequencing technologies have led to an explosion in available nucleotide sequence data, greatly enhancing our understanding of the genomic basis of many biological processes. However, the short length of the raw reads means high coverage is required for reliable sequence assembly. Nanopore sequencing has emerged as a candidate to supercede present generation sequencing and allow for theoretically unlimited read length (1). A number of strategies have been proposed, with the common basis of detecting individual nucleotides as they pass through a nanometer-scale aperture in a thin membrane separating two electrolytes (2). To date, a significant obstacle of nanopore approaches has been overcoming the fast stochastic motion of the individual molecules as they are driven through the pore (3,4). Ideally, passage of DNA through the pore would be unidirectional and each base would have a well-resolved signal. Recent methods have demonstrated an ability to controllably ratchet the DNA molecules through a nanopore one base at a time, although motion of the molecule can still occur in both forward and backward directions within a single read (5–7). Unidirectional motion remains difficult to reliably achieve, leading to a source of error in the read sequence recognized, but not previously addressed, by existing models.

In this Letter, we analyze the effect of bidirectional motion on read accuracy and propose a statistical method to account for this error. The method uses hidden Markov models (HMMs), which have been used to study multibase resolution in a nanopore sequencer (8), but have not been applied to the problem of diffusive motion inside the pore. We show that combining multiple reads from an input sequence allows accurate sequence inference, both in the presence of highly diffusive molecular motion and high base-call error rates.

Assuming no DNA-pore interaction, polymer translocation is modeled as one-dimensional diffusion with drift, with probability of displacement x in time interval t given by

$$p(x,t) = \frac{1}{\sqrt{4\pi Dt}}\exp\left[-\frac{(x-vt)^2}{4Dt}\right]. \quad (1)$$

The drift velocity $v = F/\gamma$ is determined by the driving force F and drag coefficient γ, which also determine the diffusion constant $D = \gamma k_B T$ via the fluctuation-dissipation theorem. A nondimensional forward bias is defined as $\xi = Fa/4kT$. We assume F is tuned to obtain an expected displacement $v\tau = a$, where a is one nucleotide distance and τ is the sampling interval. Defining a discretized sequence position $z = \mathrm{nint}(x/a)$, the probability of moving to position z_s given a previous position z_{s-1}, is given by

$$p(z_s \mid z_{s-1}) = \int_{a(z_s - z_{s-1})-a/2}^{a(z_s - z_{s-1})+a/2} p(x,\tau)dx. \quad (2)$$

Given an input DNA sequence of length L, we generate an output read by stepping through the sequence with discrete transition probabilities $p(z_s|z_{s-1})$, at each step making a base-call with error probability ϵ, which is independent of errors due to backward motion. We assume an appropriate method of making a base-call from the raw signal, which has been studied in O'Donnell et al. (9) in a forward-motion

Editor: David Rueda.

http://dx.doi.org/10.1016/j.bpj.2015.03.013

only model, and in Laszlo et al. (10) in a quadromer-mapping model. Next, we generate a set of N reads X_n, each having a unique length T_n. Multiple reads could arise from PCR amplification or single-molecule resequencing. Polymerase fidelity can introduce errors during amplification. We absorb this into base-call error ϵ. A schematic nanopore sequencing device is shown in Fig. 1 *A*. The relationships among forward bias, drag coefficient, and bandwidth required for single-base resolution is shown in Fig. 1 *B*. In Fig. 1 *C* we plot $p(z_s|z_{s-1})$ at different forward biases. Fig. 1 *D* shows the distribution of read lengths.

Given the set of N read sequences, the statistical task is to infer the sequence most likely to have generated the observed data. An experiment similar to this model was demonstrated on very short sequences with tunneling current data in Ohshiro et al. (11). Here, we extend the approach to the longer sequences expected from a nanopore device.

In our HMM formulation, each output read is modeled as a discrete set of observed states, $\mathbf{x} = \{x_1,\ldots x_T\}$, $x_i \in (A,C,G,T)$, which is a vector of observed bases; and a discrete set of hidden states, $\mathbf{z} = \{z_1,\ldots z_T\}$, $z_i \in (1\ldots L)$, which is the unknown position along the sequence. An HMM is described by three model parameters: the initial state distribution $\pi = p(\mathbf{z}_t)$, the hidden state transition matrix $A = p(\mathbf{z}_t|\mathbf{z}_{t-1})$, and an emission distribution $S = p(\mathbf{x}_n|\mathbf{z}_n)$ (12). The values π and A are fixed by the experimental conditions. The elements of A are obtained by numerically integrating Eq. 2 over possible transitions, δ. The inference problem in this model is to estimate the emission distribution, S, which acts as an implicit representation of our sequence,

$$S_{dl} = (1-\epsilon)p(\mathbf{x}_n = d \mid \mathbf{z}_n = l) + \epsilon/4. \quad (3)$$

In practice, S is a $4 \times L$ matrix with a multinomial distribution over the possible nucleotides at each position (Fig. 2). We use the expectation-maximization algorithm to maximize the likelihood, $p(X|\theta)$, with respect to the model parameters (13,14). The joint probability of data and states can be written as a product over the independent output, and reads

$$p(\mathbf{X},\mathbf{Z} \mid \theta) = \prod_n p(\mathbf{X}_n,\mathbf{Z}_n \mid \theta),$$

from which follows that we can perform expectation updates on each read individually before averaging results in the maximization step (see the Supporting Material for full model derivation). The resulting shared parameter estimation scheme incorporates all reads while allowing an efficient, parallel calculation.

After satisfying a convergence criterion on the likelihood ($\Delta LL < 10^{-5}$), we recover an estimated emission distribution S, which we convert to an estimate for the DNA sequence by taking $\max_d S_{dl}$. The final inference accuracy is measured as the Levenshtein distance between the input sequence and inferred sequence, normalized by L. The algorithm has complexity $O(NLT)$, where T is $O(L)$ in the limit of $\xi \to \infty$. A run of 100 reads of $L = 1$ kb DNA completes in <5 min. An example of the output of this algorithm showing the relationship between the true sequence and the inferred sequence distribution is shown in Fig. 2.

We examined the performance of the algorithm over a range of parameter values to identify a minimal experimental configuration capable of sequence inference at a given accuracy. First, we consider how to select realistic values for our parameters. $\xi = 0$ corresponds to unbiased diffusion and is unlikely to reach the end of the sequence without first exiting from the *cis* side of the nanopore (these reads are discarded in the data generation). $\xi \to \infty$ corresponds to nondiffusive motion and is a trivial case in this model. Lu et al. (4) report an experimental bias of $\xi = 0.2$ with $\gamma = 1.27$, for which they show single-pass accurate sequence recovery is impossible. We take this as a starting point for investigation, exploring the range $\xi = 0.2$ to $\xi = 10$ (from $f \approx 0.22$ MHz to $f \approx 11.2$ MHz at $\gamma = 127$). An appropriate base-call error rate will be device-specific and is

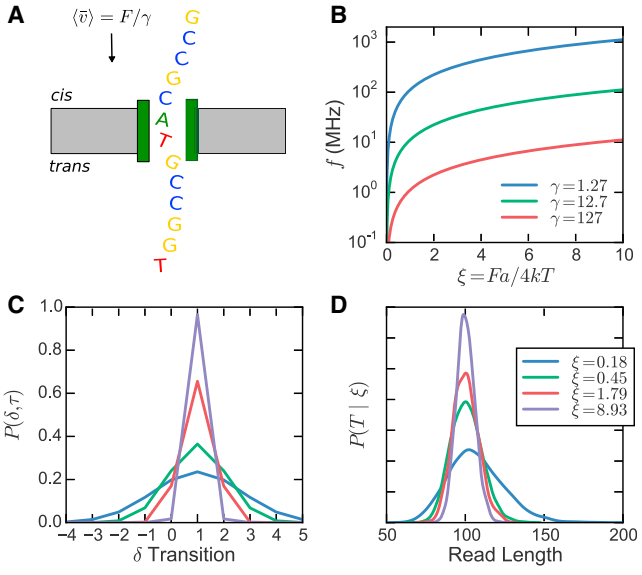

FIGURE 1 Random walk model of nanopore sequencing. (*A*) Schematic representation of an idealized nanopore sequencing device. (*B*) Minimum bandwidth required for single-base resolution as a function of forward bias, ξ, and drag coefficient, γ, given by $f = 4kT\xi/a^2\gamma$. (*C*) Transition probabilities and (*D*) read-length distributions at different forward biases (sequence length $L = 100$). To see this figure in color, go online.

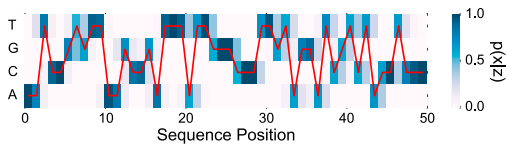

FIGURE 2 An output sequence inference distribution, *S*. Taking argmax of each column yields the called sequence (*red*). To see this figure in color, go online.

presently unknown, so we examine the range $\epsilon = 0$–0.5. The number of reads, N, is examined until convergence.

A sweep across the parameter space is shown in Fig. 3. In each sweep, the error rate was set to $\epsilon = 0.05$ unless otherwise specified. In Fig. 3 *A* we vary the forward bias for a fixed number of reads. The strongest determinant of inference accuracy is the forward bias, controlling how diffusive the motion is. However, the effect of multiple reads is immediately apparent, improving the accuracy markedly until beginning to saturate above $N = 100$. This is expected because each read contributes an independent observation of the input DNA sequence. Accuracy saturates at $\sim\xi = 1.5$—impressive, given that the probability of a single-base forward transition, $p(\delta = +1)$, is only 0.6. In Fig. 3 *B* we plot the relationship between forward bias and the number of reads in the region of $\xi = 1.5$. Accuracy improves with an increasing number of reads until convergence. Finally, inference accuracy is robust to base-call error rates up to 25% (Fig. 3, *C* and *D*). We conclude a threshold accuracy of 99% is achievable for $\xi > 2$ at an error rate $\epsilon = 0.2$. Larkin et al. (15) report a rate of $\gamma \approx 100$, corresponding to a minimum bandwidth of 2 MHz, near present measurement bandwidths. Previous estimates have placed this threshold in the GHz range; our results suggest bandwidths three orders-of-magnitude lower.

Unidirectional motion of the DNA sequence inside the nanopore has been assumed a prerequisite for accurate nanopore sequencing. We have demonstrated that accurate inference is possible, even in the presence of diffusive motion, by modeling data under an appropriate statistical model. The model combines multiple reads of the input sequence to yield a joint estimate of the true sequence. More complex translocation dynamics can be modeled by modifying the state transition matrix. Accurate sequence inference is achievable with modest improvements to the experimental constraints of present nanopore devices ($\xi \approx 0.2$). These results suggest that while amplification is necessary for high accuracy, bidirectional motion is not a limiting step in designing a nanopore sequencing device.

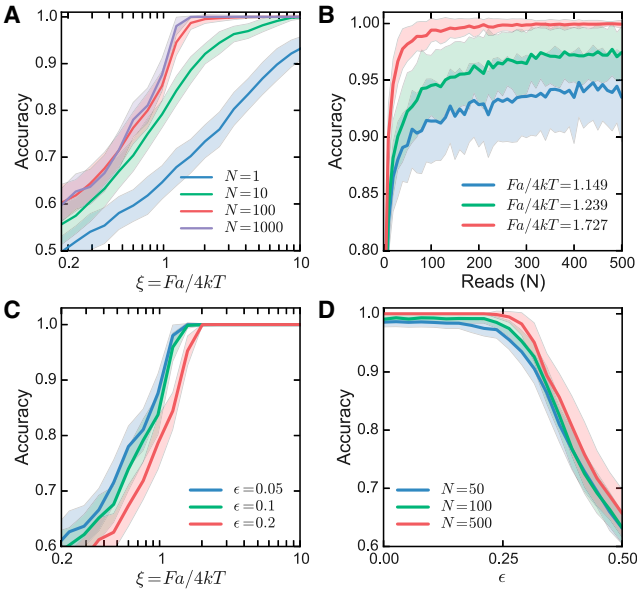

FIGURE 3 Parameter sweeps. Inference accuracy (measured as 1 - Levenshtein distance/length) is plotted as a function of experimental parameters; $\epsilon = 0.05$, unless otherwise specified. (*A*) Sweep across forward bias, β. (*B*) Sweep across number of reads, *N*. (*C*) Asymptotic performance (large *N*) performance versus forward bias β. (*D*) Sweep across error rate, ϵ. To see this figure in color, go online.

SUPPORTING MATERIAL

Model Derivation is available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(15)00273-8.

AUTHOR CONTRIBUTIONS

K.J.E., J.K.R., C.H.W., and K.L.S. designed research; K.J.E. and J.K.R. performed experiments and analyzed results; and K.J.E., J.K.R., J.-W.v.d.M., C.H.W., and K.L.S. contributed to the article.

ACKNOWLEDGMENTS

The authors gratefully acknowledge helpful discussions with David Blei, Frank Wood, David Pfau, and Peter Sims.

K.J.E. and C.H.W. were supported by National Institutes of Health grant No. U54-CA121852 (National Center for Multiscale Analysis of Genomic and Cellular Networks). J.-W.v.d.M. was supported through the Netherlands Organisation for Scientific Research Rubicon Fellowship No. 680-50-1016. K.L.S. was supported by National Institutes of Health grant No. R01-HG006879.

REFERENCES

1. Branton, D., D. W. Deamer, …, J. A. Schloss. 2008. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* 26:1146–1153.
2. Winters-Hilt, S., and M. Akeson. 2004. Nanopore cheminformatics. *DNA Cell Biol.* 23:675–683.
3. Venkatesan, B. M., and R. Bashir. 2011. Nanopore sensors for nucleic acid analysis. *Nat. Nanotechnol.* 6:615–624.
4. Lu, B., F. Albertorio, …, J. A. Golovchenko. 2011. Origins and consequences of velocity fluctuations during DNA passage through a nanopore. *Biophys. J.* 101:70–79.
5. Luan, B., H. Peng, …, G. Martyna. 2010. Base-by-base ratcheting of single stranded DNA through a solid-state nanopore. *Phys. Rev. Lett.* 104:238103.
6. Olasagasti, F., K. R. Lieberman, …, M. Akeson. 2010. Replication of individual DNA molecules under electronic control using a protein nanopore. *Nat. Nanotechnol.* 5:798–806.
7. Cherf, G. M., K. R. Lieberman, …, M. Akeson. 2012. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat. Biotechnol.* 30:344–348.
8. Timp, W., J. Comer, and A. Aksimentiev. 2012. DNA base-calling from a nanopore using a Viterbi algorithm. *Biophys. J.* 102:L37–L39.

9. O'Donnell, C. R., H. Wang, and W. B. Dunbar. 2013. Error analysis of idealized nanopore sequencing. *Electrophoresis.* 34:2137–2144.

10. Laszlo, A. H., I. M. Derrington, …, J. H. Gundlach. 2014. Decoding long nanopore sequencing reads of natural DNA. *Nat. Biotechnol.* 32:829–833.

11. Ohshiro, T., K. Matsubara, …, T. Kawai. 2012. Single-molecule electrical random resequencing of DNA and RNA. *Sci. Rep.* 2:501.

12. Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE.* 77:257–286.

13. Baum, L. E., T. Petrie, …, N. Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* 41:164–171.

14. Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B. Met.* 39:1–38.

15. Larkin, J., R. Henley, …, M. Wanunu. 2013. Slow DNA transport through nanopores in hafnium oxide membranes. *ACS Nano.* 7:10121–10128.